



Probabilistic Signal Recovery and Random Matrices

**Roman Vershynin
UNIVERSITY OF MICHIGAN**

**12/08/2016
Final Report**

DISTRIBUTION A: Distribution approved for public release.

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Executive Services, Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.</p>				
1. REPORT DATE (DD-MM-YYYY) 14-12-2016		2. REPORT TYPE Final Performance		3. DATES COVERED (From - To) 01 Jan 2014 to 31 Dec 2016
4. TITLE AND SUBTITLE Probabilistic Signal Recovery and Random Matrices			5a. CONTRACT NUMBER	
			5b. GRANT NUMBER FA9550-14-1-0009	
			5c. PROGRAM ELEMENT NUMBER 61102F	
6. AUTHOR(S) Roman Vershynin			5d. PROJECT NUMBER	
			5e. TASK NUMBER	
			5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) UNIVERSITY OF MICHIGAN 503 THOMPSON ST ANN ARBOR, MI 48109-1340 US			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AF Office of Scientific Research 875 N. Randolph St. Room 3112 Arlington, VA 22203			10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/AFOSR RTA2	
			11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-AFOSR-VA-TR-2016-0369	
12. DISTRIBUTION/AVAILABILITY STATEMENT A DISTRIBUTION UNLIMITED: PB Public Release				
13. SUPPLEMENTARY NOTES				
14. ABSTRACT <p>Our research program spanned several areas of mathematics and data science. In the area of highdimensional inference, we showed that classical methods for linear regression (such as Lasso) are applicable for non-linear data. This surprising finding has already found several applications in the analysis of genetic, fMRI and proteomic data, compressed sensing, coding and quantization. In the area of network analysis, we showed how to detect communities in sparse networks by using semidefinite programming and regularized spectral clustering. In high dimensional convex geometry, we studied the complexity of convex sets. In numerical linear algebra, we analyzed the fastest known randomized approximation algorithm for computing the permanents of matrices with non-negative entries. In computational graph theory, we studied a randomized algorithm for estimating the number of perfect matchings in general graphs. In random matrix theory, we established delocalization of eigenvectors for a wide class of random matrices, proved a sharp invertibility result for sparse random matrices, showed how to improve the norm of a general random matrix by removing a small submatrix, and developed a simple and general tool for bounding the deviation of random matrices on arbitrary geometric sets. This has applications for dimension reduction, regression and compressed sensing.</p>				
15. SUBJECT TERMS <p>high-dimensional data, Compressive Sensing, data privacy, randomization, data structures, statistical estimators</p>				
16. SECURITY CLASSIFICATION OF:				

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

DISTRIBUTION A: Distribution approved for public release.

a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified	17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON RIECKEN, RICHARD
					19b. TELEPHONE NUMBER <i>(Include area code)</i> 703-941-1100

Final Report for FA9550-14-1-0009: Probabilistic Signal Recovery and Random Matrices

Our research program supported by this grant spanned several areas of mathematics and data science. It resulted in significant discoveries in high-dimensional inference and high-dimensional probability and lead to a variety of applications in statistics, biomedical data analysis, quantization, dimension reduction, and networks science.

1. HIGH-DIMENSIONAL INFERENCE AND GEOMETRY

Our main and surprising discovery was how that many classical methods that were designed for structured *linear* regression provably work even for *non-linear* data [9, 13, 21]. The non-linearity can be very general: discontinuous, not one-to-one, and even unknown. In spite of this, we showed that methods for linear regression, such as Lasso, stay unharmed even in presence of such nonlinearities. This dramatically extends the range of statistical models for which data analysis can be done rigorously. Our findings have found a variety of applications for quantization and compressed sensing [20], as well as in the analysis of biomedical data [3, 4].

As an example, our results apply for *binary*, 0/1 measurements, which arise e.g. in classification problems and quantization. For such measurements, we also expanded the range of probability distributions the non-linear recovery results apply for. Our original analysis for non-linear data [11, 12] was done under the premise of *gaussian* measurements. In the new work [1], we showed extended it to *general nonlinear sub-gaussian* measurements.

In the area of discrete and computational geometry, we analyzed how many random hyperplanes are needed to cut a given set K in \mathbb{R}^n into much smaller pieces [10]. It turned out that the optimal number of cutting hyperplanes is proportional to the single geometric parameter of the set K , namely the *effective dimension* $d(K)$. This complexity parameter is also known to govern the efficacy of algorithms in high dimensional inference and compressed sensing. In particular, the optimal number of measurements in our work on non-linear data happened to be proportional to the same parameter – the effective dimension of the feasible set K , see [21]. Through this link, our work on cutting hyperplanes has implications in quantization, coding, dimension reduction, and compressed sensing.

Another natural measure of complexity of a convex set K is the number of faces of a polytope P that approximates K to within a constant precision. To encode a high-dimensional convex body in a form allowing computer processing, one has to construct an oracle, i.e., an algorithm that using coordinate of a point as an input, outputs whether the point is contained in

the body, or not. Construction of an oracle for a general convex body is known to be computationally hard. A potentially possible way to bypass this obstacle was suggested by Barvinok. He proposed to approximate a given body by a projection of a section of a simplex in a higher dimension. This new body, being the feasible set of a linear programming problem, allows an efficient construction of the oracle. The complexity of this construction is determined by the dimension of the simplex. An approximation with a simplex of the dimension polynomial in the dimension of the original body would have allowed to circumvent the computational hardness of the oracle construction. In the previous work of the co-PI in collaboration with A. Litvak and N. Tomczak-Jaegermann showed that, in general, such approximation is impossible. This, however, left open a possibility of construction such approximation for some important classes of convex bodies, primarily, for convex bodies with coordinate symmetries. Nevertheless, we showed in [15] that even such highly symmetric convex bodies require a simplex of exponential dimension to produce such approximation, making bypassing the hardness obstacle impossible.

2. NETWORKS

In the area of network analysis, we developed and rigorously analyzed algorithmic methods for finding structure in sparse networks [6, 7, 5]. There had been an abundance of algorithmic methods for data mining in relatively dense networks, where an average vertex has degree $\gtrsim \log n$, i.e. is connected to at least $\gtrsim \log n$ other vertices or so. Most of these methods, including the most popular Principal Component Analysis (PCA), manifestly fail for sparser networks, in particular for those with *constant* average degrees.

Practitioners had suggested that the problem for sparse networks lies in the vertices of abnormally high degrees, and suggested that regularizing those vertices by pruning or lowering their weight could solve the problem. We confirmed this rigorously by showing a very general result: any regularization pre-processing which brings the degrees down to normal provably, leads to spectral concentration, and therefore makes spectral methods like PCA work [7].

In a related development [5], we proved for the first time that methods based on semidefinite programming also work for structure discovery in sparse networks. Our analysis is based on Grothendieck's inequality. In all previous applications in theoretical computer science had only yielded approximation to within some fixed constant factor. We demonstrated a new method where Grothendieck's inequality can be used to give an arbitrarily fine approximation.

For both methods, our theory is applicable for a far wider class of networks than the benchmark class of stochastic block models that is usually discussed in network science results.

3. PERMANENTS, HAFNIANS AND PERFECT MATCHINGS

In numerical linear algebra, we studied the fastest known randomized approximation algorithm for computing the *permanents of matrices* with non-negative entries, namely the Barvinok-Godsil-Gutman estimator. The permanent is an important computational characteristic which counts, for instance, the number of perfect matchings in a bipartite graph. Besides this, permanents arise naturally in the study of contingency tables, evaluation of the expected product of dependent normal random variables, etc. It is known that the evaluation of a permanent is a $\#$ -P hard problem, so taking into account the limitations of the computing power, one can hope only to estimate it. Barvinok-Godsil-Gutman estimator probabilistic estimator is the fastest known means of estimating the permanent. In the worst-case scenario, it outputs the permanent with the multiplicative error which is exponential in the size of the matrix. Yet, it has been observed that, typically, the actual performance of this estimator is much better than the worst case. We discovered a sufficient condition on a deterministic graph or matrix guaranteeing a smaller error for estimating the permanent [19].

In a related development in computational graph theory [16], we analyzed a probabilistic algorithm for estimating the number of perfect matchings in general graphs. Unlike bipartite graphs, where the number of perfect matchings is represented by the permanent of the adjacency matrix, in a general case, it is evaluated by a much more complex quantity, namely the *hafnian* of the same matrix. Because of this additional complexity, almost all known methods of estimating the number of perfect matchings which were developed for bipartite graphs fail for the general ones. The only exception is the Barvinok estimator which is currently the unique polynomial time probabilistic estimator for the number of perfect matchings. This fact makes the error analysis for this estimator especially important. As for bipartite graphs, the worst case error of this estimator is exponential in the size of the graph. We showed that if the graph possesses certain expansion properties, then the error of the Barvinok estimator is much smaller than in the worst case.

4. RANDOM MATRIX THEORY

Several significant advances were made in random matrix theory and its applications. We established *delocalization of eigenvectors* for a wide class of random matrices [17, 18]. This means that with high probability, every eigenvector of a random matrix is delocalized in the sense that any subset of

its coordinates carries a non-negligible portion of its L^2 norm. Our results pertain to a wide class of random matrices, including matrices with independent entries, symmetric and skew-symmetric matrices, as well as some other naturally arising ensembles.

Next, we analyzed in [2] the *condition number* of sparse random matrices, a quantity important for controlling the running time and the precision of various numerical linear algebra algorithms. This is an important problem especially for sparse random matrix, which are among the basic tools in statistics, computer sciences and signal processing. As we increase sparsity, we found that the condition number stays nearly the same as for a dense matrix almost until the transition point where an entire zero row appears (at which point it obviously jumps to infinity).

Furthermore, we showed how to improve the behavior of a random matrix by modifying a small fraction of its entries [14]. We studied the conditions where the operator norm of a random matrix A can be reduced to the optimal order by zeroing out a small submatrix of A . We found that this is possible if and only if the entries of A have zero mean and finite variance. Moreover, we obtained an almost optimal dependence between the size of the removed submatrix and the resulting operator norm.

Finally, we developed a simple and general tool for bounding the deviation of random matrices on arbitrary geometric sets [8]. This new deviation inequality unified many existing results, such as Johnson-Lindenstrauss Lemma which plays a major role in dimension reduction, M^* bound in high-dimensional convex geometry, and a non-asymptotic version of Bai-Yin limiting law in random matrix theory. On top of that, our deviation inequality led to several new applications, in particular for dimension reduction, model selection, structured regression and compressed sensing [8].

REFERENCES

- [1] A. Ai, A. Lapanowski, Y. Plan and R. Vershynin, *One-bit compressed sensing with non-Gaussian measurements*, 2015 IEEE International Conference on Bioinformatics and Biomedicine, 994-998.
- [2] A. Basak, M. Rudelson, *Invertibility of sparse non-hermitian matrices*, submitted.
- [3] S. Chretien, C. Gueux, M. Boyer-Guittaut, R. Delage-Mouroux, F. Descotes, *Investigating gene expression array with outliers and missing data in bladder cancer*, preprint.
- [4] M. Genzel, G. Kutyniok, *A mathematical framework for feature selection from real-world data with non-linear observations*, preprint.
- [5] O. Guedon, R. Vershynin, *Community detection in sparse networks via Grothendieck's inequality*, Probability Theory and Related Fields 165 (2016), 1025–1049.

- [6] C. Le, E. Levina, R. Vershynin, *Optimization via low-rank approximation, with applications to community detection in networks*, Annals of Statistics 44 (2016), 373–400.
- [7] C. Le, E. Levina, R. Vershynin, *Concentration and regularization of random graphs*, Random Structures and Algorithms, to appear.
- [8] C. Liaw, A. Mehrabian, Y. Plan, R. Vershynin, *A simple tool for bounding the deviation of random matrices on geometric sets*, Geometric Aspects of Functional Analysis, Lecture Notes in Mathematics, Springer, Berlin, to appear.
- [9] Y. Plan, R. Vershynin, *The generalized Lasso with non-linear observations*, IEEE Transactions on Information Theory, to appear.
- [10] Y. Plan, R. Vershynin, *Dimension reduction by random hyperplane tessellations*, Discrete and Computational Geometry 51 (2014), 438–461.
- [11] Y. Plan, R. Vershynin, *One-bit compressed sensing by linear programming*, Communications on Pure and Applied Mathematics 66 (2013), 1275–1297.
- [12] Y. Plan, R. Vershynin, *Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach*, IEEE Transactions on Information Theory 59 (2013), 482–494.
- [13] Y. Plan, R. Vershynin, E. Yudovina, *High-dimensional estimation with geometric constraints*, Information and Inference 0 (2016), 1–40.
- [14] E. Rebrova, R. Vershynin, *Norms of random matrices: local and global problems*, submitted.
- [15] M. Rudelson, *On the complexity of the set of unconditional convex bodies*, Discrete Comput. Geom. 55 (2016), no. 1, 185–202.
- [16] M. Rudelson, A. Samorodnitsky, O. Zeitouni, *Hafnians, perfect matchings and Gaussian matrices*, Annals of Probability, to appear.
- [17] M. Rudelson, R. Vershynin, *Delocalization of eigenvectors of random matrices with independent entries*, Duke Mathematical Journal 164 (2015), 2507–2538.
- [18] M. Rudelson, R. Vershynin, *No-gaps delocalization for general random matrices*, Geometric and Functional Analysis, to appear.
- [19] M. Rudelson, O. Zeitouni, *Singular values of gaussian matrices and permanent estimators*, Random Structures Algorithms 48 (2016), 183–212.
- [20] H.-J. Shi, M. Case, X. Gu, S. Tu, D. Needell, *Methods for Quantized Compressed Sensing*, preprint.
- [21] R. Vershynin, *Estimation in high dimensions: a geometric perspective*. Sampling Theory, a Renaissance, 3–66, Birkhauser Basel, 2015.

AFOSR Deliverables Submission Survey

Response ID:7238 Data

1.

Report Type

Final Report

Primary Contact Email

Contact email if there is a problem with the report.

romanv@umich.edu

Primary Contact Phone Number

Contact phone number if there is a problem with the report

7349294092

Organization / Institution name

University of Michigan

Grant/Contract Title

The full title of the funded effort.

Probabilistic Signal Recovery and Random Matrices

Grant/Contract Number

AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-14-1-0009

Principal Investigator Name

The full name of the principal investigator on the grant or contract.

Roman Vershynin

Program Officer

The AFOSR Program Officer currently assigned to the award

Richard Riecken

Reporting Period Start Date

01/01/2014

Reporting Period End Date

12/31/2016

Abstract

Our research program spanned several areas of mathematics and data science. In the area of high-dimensional inference, we showed that classical methods for linear regression (such as Lasso) are applicable for non-linear data. This surprising finding has already found several applications in the analysis of genetic, fMRI and proteomic data, compressed sensing, coding and quantization. In the area of network analysis, we showed how to detect communities in sparse networks by using semidefinite programming and regularized spectral clustering. In high dimensional convex geometry, we studied the complexity of convex sets. In numerical linear algebra, we analyzed the fastest known randomized approximation algorithm for computing the permanents of matrices with non-negative entries. In computational graph theory, we studied a randomized algorithm for estimating the number of perfect matchings in general graphs. In random matrix theory, we established delocalization of eigenvectors for a wide class of random matrices, proved a sharp invertibility result for sparse random matrices, showed how to improve the norm of a general random matrix by removing a small submatrix, and developed a simple and general tool for bounding the deviation of random matrices on arbitrary geometric sets. This has applications for dimension reduction, regression and compressed sensing.

DISTRIBUTION A: Distribution approved for public release.

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your [SF298](#) form. A blank SF298 can be found [here](#). Please do not password protect or secure the PDF. The maximum file size for an SF298 is 50MB.

[sf0298.pdf](#)

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF. The maximum file size for the Report Document is 50MB.

[AF-Report-final.pdf](#)

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

A. Ai, A. Lapanowski, Y. Plan and R. Vershynin, One-bit compressed sensing with non-Gaussian measurements, 2015 IEEE International Conference on Bioinformatics and Biomedicine, 994-998.

A. Basak, M. Rudelson, Invertibility of sparse non-hermitian matrices, submitted.

M. Genzel, G. Kutyniok, A mathematical framework for feature selection from real-world data with non-linear observations, preprint.

O. Guedon, R. Vershynin, Community detection in sparse networks via Grothendieck's inequality, Probability Theory and Related Fields 165 (2016), 1025–1049.

C. Le, E. Levina, R. Vershynin, Optimization via low-rank approximation, with applications to community detection in networks, Annals of Statistics 44 (2016), 373–400.

C. Le, E. Levina, R. Vershynin, Concentration and regularization of random graphs, Random Structures and Algorithms, to appear.

C. Liaw, A. Mehrabian, Y. Plan, R. Vershynin, A simple tool for bounding the deviation of random matrices on geometric sets, Geometric Aspects of Functional Analysis, Lecture Notes in Mathematics, Springer, Berlin, to appear.

Y. Plan, R. Vershynin, The generalized Lasso with non-linear observations, IEEE Transactions on Information Theory, to appear.

Y. Plan, R. Vershynin, Dimension reduction by random hyperplane tessellations, Discrete and Computational Geometry 51 (2014), 438-461.

Y. Plan, R. Vershynin, One-bit compressed sensing by linear programming, Communications on Pure and Applied Mathematics 66 (2013), 1275–1297.

Y. Plan, R. Vershynin, Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach, IEEE Transactions on Information Theory 59 (2013), 482–494.

Y. Plan, R. Vershynin, E. Yudovina, High-dimensional estimation with geometric constraints, Information and Inference 0 (2016), 1–40.

E. Rebrova, R. Vershynin, Norms of random matrices: local and global problems, submitted.

M. Rudelson, On the complexity of the set of unconditional convex bodies, Discrete Comput. Geom. 55 (2016), 185–202.

M. Rudelson, A. Samorodnitsky, O. Zeitouni, Hafnians, perfect matchings and Gaussian matrices, Annals of Probability, to appear.

M. Rudelson, R. Vershynin, Delocalization of eigenvectors of random matrices with independent entries, Duke Mathematical Journal 164 (2015), 2507–2538.

M. Rudelson, R. Vershynin, No-gaps delocalization for general random matrices, Geometric and Functional Analysis, to appear.

M. Rudelson, O. Zeitouni, Singular values of gaussian matrices and permanent estimators, Random Structures Algorithms 48 (2016), 183–212.

R. Vershynin, Estimation in high dimensions: a geometric perspective. Sampling Theory, a Renaissance, 3–66, Birkhauser Basel, 2015.

New discoveries, inventions, or patent disclosures:
Do you have any discoveries, inventions, or patent disclosures to report for this period?
 No

Please describe and include any notable dates
Do you plan to pursue a claim for personal or organizational intellectual property?
Changes in research objectives (if any):

Change in AFOSR Program Officer, if any:
 Dr. Tristan Nguyen to Dr. Richard Riecken

Extensions granted or milestones slipped, if any:

AFOSR LRIR Number
LRIR Title

Reporting Period
Laboratory Task Manager
Program Officer

Research Objectives
Technical Summary

Funding Summary by Cost Category (by FY, \$K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document
Report Document - Text Analysis
Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

Nov 10, 2016 09:28:01 Success: Email Sent to: romanv@umich.edu